DOCUMENT RESUME

ED 174 103 HE 011 444

AUTHOR Marshall, K. T.; Oliver, R. M.
TITLE Estimating Errors in Student Enrollment Forecasting. AIR Forum 1979 Paper.
PUB DATE May 79
NOTE 22p.; Paper presented at the Annual Forum of the Association for Institutional Research (19th, San Diego, California, May 13-17, 1979)

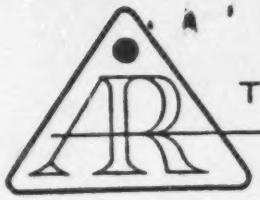
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *College Attendance; *Enrollment Projections; Higher Education; Institutional Research; Longitudinal Studies; Mathematical Formulas; *Mathematical Models; *Prediction; Probability Theory; *Research Methodology; *Statistical Analysis; Validity
IDENTIFIERS AIR Forum 1979

ABSTRACT

The use of data on longitudinal student attendance patterns to determine variances, and hence confidence bounds, on student enrollment forecasts, in addition to finding the forecasts themselves, is demonstrated. The formulation of the enrollment model based on longitudinal student attendance patterns is described step by step, presenting the equations used in the model and some sample data obtained through application of the model. Applications of various equations are demonstrated. The model is analyzed assuming that the new admissions each period follow a Poisson distribution. The enrollment permits the development of statements about the approximate behavior of the conditional distribution of the total number of students in attendance at a given time. It is assumed that entering cohorts contain large numbers of students. Equations relative to this function are presented and discussed. Sample data on the forecasts and confidence intervals for each semester enrollment are presented. The model provides a measure of the error that could be present for each forecast. (SC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED174103



THE ASSOCIATION FOR INSTITUTIONAL RESEARCH

This paper was presented at the Nineteenth Annual Forum of the Association for Institutional Research held at the Town and Country Inn in San Diego, California, May 13-17, 1979. This paper was reviewed by the AIR Forum Publications Committee and was judged to be of high quality and of interest to others concerned with the research of higher education. It has therefore been selected to be included in the ERIC collection of Forum papers.

Mary Corcoran
University of Minnesota
(Editor, AIR Forum Publications)

NE 011 444

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
AIR
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ESTIMATING ERRORS IN STUDENT ENROLLMENT FORECASTING

[†]K. T. Marshall

and

[††]R. M. Oliver

[†]Naval Postgraduate School, Monterey, Ca.

[††]University of California, Berkeley, Ca.

0. Introduction

In 1968 Sidney Suslow, together with his colleagues in the Office of Institutional Research at the Berkeley Campus of the University of California, completed a study (Suslow et al. [4]) of undergraduate student attendance patterns over time. That report contains some of the earliest data the authors had seen on a given group, or cohort, of students, and how the group behaved over its undergraduate career. Most institutions keep only cross-sectional data obtained from enrollment statistics. It was the availability of the Suslow data that led the authors to pursue the formulation and analysis of enrollment models based on longitudinal student attendance patterns. The authors presented a constant-work model (Marshall and Oliver [2]) which explained the data quite successfully. They also, together with Suslow in [3], tried to find cross-sectional Markovian models to fit the longitudinal data (this latter work is reproduced in a shortened form in Chapter 2 of Grinold and Marshall [1], which is perhaps more accessible than [3]).

The purpose of this paper is to demonstrate how the longitudinal data can be used to determine variances, and hence confidence bounds, on student enrollment forecasts in addition to finding the forecasts themselves. Thus with each forecast we have a measure of the error that could be present.

1. Model Formulation

We consider discrete points in time such as the beginning of a quarter, semester, or academic year. The particular choice depends on the model use and the availability of data. In our numerical examples we use the data from Suslow et al. [4], and hence our time points coincide with semesters. Thus when we write $t = 1,2,3,\ldots$, we mean the start of the first, second, third, etc. semesters in the future; $t = 0$ will refer to the point "now" from which forecasts are being made, and $t = -1, -2, -3$, will refer to the first, second, third, etc. semesters in the past.

Our first aim is to derive an expression for the expected number in attendance at some time $t > 0$. We do not differentiate groups such as freshmen, sophomores, or lower division, upper division. This could easily be done by placing subscripts on our notation, but we choose to simplify the notation to be consistent with the Suslow data on total student attendance.

Let $S(t;u)$ be the number of students in attendance at time t who entered (for the first time) at time $t - u$, $u = 0,1,\ldots$. Let $S(t)$ be the total number of students in attendance at time t. Then

$$S(t) = S(t;0) + S(t;1) + S(t;2) + \cdots + S(t;u) + \cdots . \qquad (1)$$

The data in [4] showed that for the periods studied (1950's and 1960's) there was very stable behavior in student attendance; the fraction of students who attended a given semester

after entrance was independent of when the students first entered. However, only fall-entering cohorts were studied. We assume here that stable behavior could be expected from spring-entering cohorts also, but that fall- and spring-entering students could have different continuation fractions. Let $p_1(u)$ be the probability that a student attends at time u after entering in the fall, independent of the particular entrance time. Let $p_2(u)$ be equivalent probability for spring-entering students. We also assume that the attendance of any given student is independent of the attendance or non-attendance of any other student; i.e. all students act independently of each other. Table 1 gives $p_1(u)$ determined by Suslow et al. in [4].

Let $N(t)$ be the number of <u>new</u> students who enter at time t. The above two assumptions imply that the value of (the random variable) $S(t;u)$, given the value of $N(t-u)$, has a binomial probability distribution. That is,

$$\Pr[S(t;u) = k | N(t-u) = m] = \binom{m}{k} p_i(u)^k [1 - p_i(u)]^{m-k} , \qquad (2)$$

for $k = 0,1,\ldots,m$, and $n \geq 0$, where $i = 1$ for fall students and $i = 2$ for spring students. In particular the conditional expectation and the conditional variance of $S(t;u)$ are given respectively by

$$E[S(t;u) | N(t-u) = m] = mp_i(u) , \qquad (3)$$

$$\mathrm{Var}[S(t;u) | N(t-u) = m] = mp_i(u)[1 - p_i(u)] . \qquad (4)$$

u	$p_1(u)$	$p_1(u)(1-p_1(u))$	$p_1(u)^2$
0	1.0	0.0	1.0
1	.972	.0272	.9448
2	.905	.0860	.8190
3	.756	.1845	.5715
4	.684	.2161	.4679
5	.593	.2414	.3516
6	.562	.2462	.3158
7	.524	.2494	.2746
8	.498	.2500	.2480
9	.199	.1594	.0396
10	.130	.1131	.0169
11	.050	.0475	.0025
12	.036	.0347	.0013
13	.017	.0167	.0003
14	.015	.0148	.0002
15	.011	.0109	.0001
16	.007	.0070	.0000
	6.959	1.905	5.054

TABLE 1: Sample student attendance data from Suslow et al. [4].

Let t be the start of a fall semester. After taking expectations in (1) and using (3), the expected total enrollment at time t is

$$E[S(t)] = \sum_{u=0}^{\infty} p_{i(u)}(u)\ E[N(t-u)] \ . \qquad (5)$$

Here we have let

$$i(u) = 1 \quad \text{if} \quad u = 0,2,4,6, \ \ldots$$
$$= 2 \quad \text{if} \quad u = 1,3,5,7, \ \ldots \ .$$

For any two random variables X and Y the expression

$$Var[X] = E[Var[X|Y]] + Var[E[X|Y]]$$

holds. We use this together with (1), (3) and (4) to obtain for the variance of the total enrollment at time t,

$$Var[S(t)] = \sum_{u=0}^{\infty} \Big(E[N(t-u)]\ p_{i(u)}(u)(1 - p_{i(i)}(u)) + p_{i(u)}(u)^2\ Var(N(t-u)) \Big) \ . \qquad (6)$$

Equations (5) and (6) give the expected enrollment and its variance at time t. Recall that t is a fall semester. For the case when t is a spring semester we use

$$i(u) = 2 \quad \text{if} \quad u = 0,2,4,6, \ \ldots$$
$$= 1 \quad \text{if} \quad u = 1,3,5,7, \ \ldots \ .$$

These expressions do not take into account the fact that we have knowledge of enrollments up to time $t = 0$ (the current time in our timing convention). In (5) we know the values of $N(0)$, $N(-1)$, $N(-2)$, etc. and thus our forecast for $t > 0$ becomes

$$E[S(t)|N(0),N(-1),\ldots] = \sum_{u=t}^{\infty} p_{i(u)}(u)\, N(t-u) + \sum_{u=0}^{t-1} p_{i(u)}(u)\, E[N(t-u)]\,, \tag{7}$$

where $i(u)$ is defined above for the particular case that t is either fall or spring. The first summation term in equation (7) gives the expected "legacy" at time t of the given inputs up to and including the current time zero. The second summation gives the expected enrollment at time t from the expected input of new students at times $1, 2, \ldots, t$.

Similarly, by using equation (6), the variance of the forecast at t, given inputs up to and including time zero, becomes

$$\begin{aligned} \mathrm{Var}[S(t)|N(0),N(-1),\ldots] &= \sum_{u=t}^{\infty} p_{i(u)}(u)\,(1 - p_{i(u)}(u))\, N(t-u) \\ &+ \sum_{u=0}^{t-1} \left(p_{i(u)}(u)(1 - p_{i(u)}(u))\, E[N(t-u)] + p_{i(u)}(u)^2\, \mathrm{Var}(N(t-u)) \right). \end{aligned} \tag{8}$$

The first summation gives the contribution to the variance from the inputs up to and including the present. The second summation

gives the contribution which will occur from future inputs. Note that this depends on the variance of the new inputs for times $1,2,\ldots,t$ as well as the variance due to returning students.

Table 1 gives data for $p_1(u)$, $u \geq 0$, obtained originally in the study for Suslow et al. [4], and reproduced on page 66 of [1]. The third and fourth columns give $p_1(u)(1-p_1(u))$ and $p_1(u)^2$ respectively. These data are required in equation (8), whereas the data in column 2 are required in equation (7).

The usual interpretation given to the second column in Table 1 is simply the fraction of attending students out of a given cohort. The third column is the variance of the $S(t;u)$ terms divided by $N(t-u)$. It is interesting to see how the conditional expectation and the variance of the number of attending students vary with the number of time periods that have elapsed since initial registration. As one might expect, the fraction of students out of a given cohort that return to attend decreases rapidly and there is a sharp drop of attendance after eight semesters. By the end of the 12th semester the fraction of attending students decreases to a number less than 4% of the original cohort. However, the conditional variance of the number returning first increases, has its maximum when seven or eight semesters have elapsed and then decreases to a negligible amount by the end of the 12th semester. About the 12th semester, the conditional expectation and variance of the number attending

are about equal; this result is not surprising, if we recall that the Poisson distribution (whose variance and mean are equal) is a good approximation to the binomial distribution when the probability $p(u)$ is small. Thus, students returning after 10 periods can be classified as "rare" events in the sense that while the probability that an individual student attends is small the original cohort is large enough so that the probability distribution of returning students is Poisson. By similar arguments one can deduce that the number who do not attend in the first few semesters is also Poisson distributed.

Consider a simple system where there is no variance in the new student input, which is a fixed amount, say n_1, in each fall semester, and a fixed amount n_2 in each spring semester. Thus $E[N(t)] = n_i$ and $Var[N(t)] = 0$ for all t where $i = 1$ for a fall semester and $i = 2$ in the spring. Using these in (7) and (8), and assuming $p_1(u) = p_2(u)$ with the data in Table 1, we obtain

$$E[S(t)] = 3.873n_1 + 3.122n_2 , \qquad Var[S(t)] = 0.968n_1 + 0.937n_2$$

for t a fall semester, and

$$E[S(t)] = 3.873n_2 + 3.122n_1, \qquad Var[S(t)] = 0.968n_2 + 0.937n_1$$

for t a spring semester. All these expressions are independent of t because of the constant input each period.

Table 2 illustrates the use of these equations for three combinations of fall and spring input totalling 4000 per year, and assuming $\{p_1(u) = p_2(u)\}$ are given in Table 1.

Semester	Input	Expected Enrollment	Variance of Enrollment
Fall	4,000	15,348	3,872
Spring	0	12,488	3,748
Fall	3,000	14,633	3,841
Spring	1,000	13,203	3,779
Fall	2,000	13,918	3,810
Spring	2,000	13,918	3,810

TABLE 2: Illustrative calculations for differing fall/spring input values.

A fairly typical use for Equations (7) and (8) is that of forecasting one period into the future. With the convention that $t = 0$ represents today (the start of a fall semester), we obtain the next period forecast

$$E[S(1)|N(0), N(-1),\ldots] = \sum_{u=1}^{\infty} p_{i(u)}(u)\, N(1-u) + E[N(1)],$$

with $i(u) = 1$ for u even, $i(u) = 2$ for u odd, and provided $p_1(0) = 1$. The first (summation) terms represents the expected number of returning students and the second term represents the expected number of new admissions. The corresponding expression for the variance of enrollments in the next period is

$$\mathrm{Var}[S(1)|N(0),N(-1),\ldots] = \sum_{u=1}^{\infty} p_{i(u)}(u)\,(1 - p_{i(u)}(u))\, N(1-u) + \mathrm{Var}[N(1)].$$

In this case where we assume all entering students in fact show up, the fluctuations are due either to the uncertainty in the count of returning students already enrolled or to the uncertainty in the new students. Thus one can obtain some idea of where new forecasting efforts should be directed. In certain institutions the dominant problem may be the uncertainties associated with returning students rather than with new students. If, for example, the past cohorts were approximately 3000 in each fall and 1000 in each spring, but the next group of entering students were Poisson with expected number and variance equal to 1000 then we would have (from Table 2)

$$\mathrm{Var}[S[1]|N(0),N(-1),\ldots] \cong 3779 + 1000 = 4779\ .$$

In this case, two standard deviations (a measure of error often used and based on Normal distribution theory) would be 138 students which is slightly larger than the value we obtain when all admissions are constant ($2 \times \sqrt{3770} = 122$ from Table 2). In other words it is possible to make various assumptions about the uncertainty of future enrollments and/or returning students and easily include them in our estimates of enrollment fluctuations.

It is unlikely that student input each period would be constant. In the next section we analyze the model assuming that new admissions follow a Poisson distribution.

2. Poisson Admissions

The number of new students who actually enroll in a given future semester is not known with certainty. A simple method of modelling this uncertainty is to assume the number of new enrollments follows a Poisson distribution. Let $n(t)$ be the expected number of new enrollments at time t. Then

$$\Pr[N(t) = m] = \frac{n(t)^m \, e^{-n(t)}}{m!}, \qquad m \geq 0 . \tag{9}$$

From equations (2) and (9) we get

$$\Pr[S(t;u)=k] = \frac{p_{i(u)}(u) \; n_{i(u)}(t-u)^k \; e^{-p_{i(u)}(u) \; n_{i(u)}(t-u)}}{k!}, \quad k \geq 0 \tag{10}$$

This shows that each random variable in (1) has a Poisson distribution, which together with our independence assumption, implies that the total enrollment at time t has a Poisson distribution at every time t, with

$$E[S(t)] = \mathrm{Var}[S(t)] = \sum_{u=0}^{\infty} p_{i(u)}(u) \; n_{i(u)}(t-u) .$$

Using our previous example, but with Poisson input instead of fixed input, with $n_1 = 3000$, $n_2 = 1000$ and $p_1(u) = p_2(u)$ as in Table 1, we get again an expected enrollment

of 14,741 each fall and 13,239 each spring, but with variances of the same values. Thus two standard deviations would be 242 each fall and 230 each spring, which show much more uncertainty in the forecasts as one would expect.

3. Large Cohort Sizes

We have already shown in equation (2) that the number of students attending out of a given entering cohort can be viewed as the result of summing successes in Bernoulli trials, where the probability of success is the probability that a student attends on a given semester. Thus, if add a finite number of such random variables to obtain the attendance at a later time period we again obtain a sum of successes in a finite number of Bernoulli trials. If the parameter $p_i(u)$ of the Binomial distribution in (2) did not change with time, then it would also be true that the sum in (1) is binomially distributed. This follows from the derivation of the distribution of the sum of successes in a finite number of Bernoulli trials, each trial having the same probability of success. Unfortunately, that is not the case; as we can easily see from Table 1 the parameter $p_i(u)$ changes rather dramatically with elapsed time since entry and the resulting distribution is obtained from the convolution of as many binomial distributions, with changing parameters, as there are terms in (1). Although explicit expressions can be found for the generating function of such distributions, algebraic expressions for the distribution itself are not simple. Fortunately, however, much can be said about the approximate behavior of the conditional distribution of $S(t)$ if we assume that entering cohorts contain large numbers of students.

The central limit theorem of probability theory states that if $S(t;u)$ is the sum of the number of successes in $n(t-u)$

trials each with success probability $p_i(u)$, then the normalized sum

$$S^*(t;u) = \frac{S(t;u) - p_i(u)\ n(t-u)}{[p_i(u)(1-p_i(u))\ n(t-u)]^{1/2}} \tag{11}$$

is approximately normally distributed. If we write

$$\Phi(a) = \frac{1}{\sqrt{2}} \int_{-\infty}^{a} e^{-y^2/2} dy$$

for the normal distribution function, then with large cohort sizes, i.e., large numbers entering at $t-u$,

$$\Pr[S^*(t;u) \leq a] \approx \Phi(a) \quad \text{independent of} \quad p_i(u) \quad \text{and} \quad t. \tag{12}$$

As long as each entering cohort is large and entering cohorts act independently of one another the sum of a finite number of terms in (1) is also approximately normal. In this case

$$\Pr[S^*(t) \leq a] \approx \Phi(a)\ , \tag{13}$$

where the normalization for $S^*(t)$ is given by

$$S^*(t) = \frac{S(t) - \sum_{u \geq 0} p_{i(u)}(u)\ n(t-u)}{\left(\sum_{u \geq 0} n(t-u)\ p_{i(u)}(u)(1-p_{i(u)}(u))\right)^{1/2}} \tag{14}$$

Table 3 gives $E[S(t;u)]$ for $u = 0,1,\ldots,12$ and $E[S(t)]$ together with 95% confidence intervals. Also tabulated is the length of the confidence intervals as a percentage of the expected values. Fall and spring semesters are shown in separate columns for clarity (again t is assumed to be a fall semester). Note how the uncertainty as a percentage of the mean increases with time enrolled, and how small the error is on the total enrolled forecast compared to the individual semesters.

Equations (13) and (14) can be used to obtain more information on the uncertainty in $S(t)$; one can estimate the probability of the enrollment exceeding any given figure, of not exceeding any given figure, or of being in any given range. Let a and b be any two numbers with $a < b$. Then for $n_1 = 3000$, $n_2 = 1000$, t a fall semester, and the data given in Table 1 with $p_1(u) = p_2(u)$, then

$$\Pr[a \leq S(t) \leq b] \approx \Phi\left(\frac{b - 14,633}{62}\right) - \Phi\left(\frac{a - 14,633}{62}\right).$$

From tables of the normal distribution we see that

$$\begin{aligned} P[S(t) \leq 14,700] &\approx 0.86 , \\ P[S(t) \geq 14,500] &\approx 0.98 , \\ P[14,500 \leq S(t) \leq 14,700] &\approx 0.84 . \end{aligned} \tag{15}$$

Time u	E[S(t;u)] and 95% Confidence Interval		Confidence Interval as % of E[S(t;u)]	
	Fall	Spring	Fall	Spring
0	3000 ∓ 0		0	
1		972 ∓ 10		2.1
2	2715 ∓ 32		2.4	
3		756 ∓ 27		7.1
4	2052 ∓ 51		5.0	
5		593 ∓ 31		10.5
6	1686 ∓ 54		6.4	
7		524 ∓ 32		12.2
8	1494 ∓ 55		7.4	
9		199 ∓ 25		25.1
10	390 ∓ 37		19.0	
11		50 ∓ 14		56.0
12	108 ∓ 20		37.0	
13		17 ∓ 8		94.1
14	45 ∓ 13		57.8	
15		11 ∓ 7		127.3
16	21 ∓ 9		85.7	
Total	14,633 ∓ 124		1.7	

TABLE 3: Forecasts and confidence intervals for each semester enrollment, $n_1 = 3000$, $n_2 = 1000$, and t a fall semester.

The Normal approximation for $S(t)$ still holds if the admissions each semester are assumed to be Poisson, since the total enrollment is the sum of independent Poisson random variables with distribution given by (10). In this case we consider

$$S^*(t) = \frac{S(t) - \sum_{u \geq 0} p_{i(u)}(u)\ n(t-u)}{\left(\sum_{n \geq 0} p_{i(u)}(u)\ n(t-u) \right)^{1/2}} \ .$$

For fall Poisson inputs with mean 3000, spring Poisson inputs with mean 1000, t a fall semester, and assuming $p_1(u) = p_2(u)$ given in Table 1, then

$$P[a \leq S(t) \leq b] \approx \Phi\left(\frac{b - 14,633}{121} \right) - \Phi\left(\frac{a - 14,633}{121} \right) \ .$$

In this case

$$\begin{aligned} P[S(t) \leq 14,700] &\approx 0.71 \ , \\ P[S(t) \geq 14,500] &\approx 0.86 \ , \\ P[14,500 \leq S(t) \leq 14,700] &\approx 0.57 \ . \end{aligned} \qquad (16)$$

A comparison of (15) and (16) shows the added uncertainty in the forecast due to randomness in the numbers of admissions.

REFERENCES

[1] Grinold, R.C., and Marshall, K.T. Manpower Planning Models, Elsevier-North Holland, New York (1977).

[2] Marshall, K.T., and Oliver, R.M. "A Constant Work Model for Student Enrollment and Attendance," Operations Research, Vol. 18, pp. 193-206 (1970).

[3] Marshall, K.T., Oliver, R.M., and Suslow, S., "Undergraduate Enrollment Attendance Patterns," Report No. 4, Administrative Studies Project in Higher Education, Office of Institutional Research, University of California, Berkeley, Ca. 94720 (1970).

[4] Suslow, S., Langlois, E., Sumariwalla, R., and Walther, C., "Student Performance and Attrition at the University of California, Berkeley," Office of Institutional Research, University of California, Berkeley, Ca. 94720 (1968).